2008-11

Using the Repeated Two-sample Rank Procedure for Detecting Anomalies in Space and Time

Fricker, Ronald D., Jr.

NAVAL
POSTGRADUATE
SCHOOL

Using the Repeated Two-sample Rank Procedure for Detecting Anomalies in Space and Time

Ronald D. Fricker, Jr.

University of California, Riverside

November 18, 2008

What is Biosurveillance?

- Homeland Security Presidential Directive HSPD-21 (October 18, 2007):
  - "The term 'biosurveillance' means the process of active data-gathering … of biosphere data … in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity." [1]
  - "The Secretary of Health and Human Services shall establish an operational national epidemiologic surveillance system for human health..." [1]
- Epidemiologic surveillance:
  - "…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." [2]

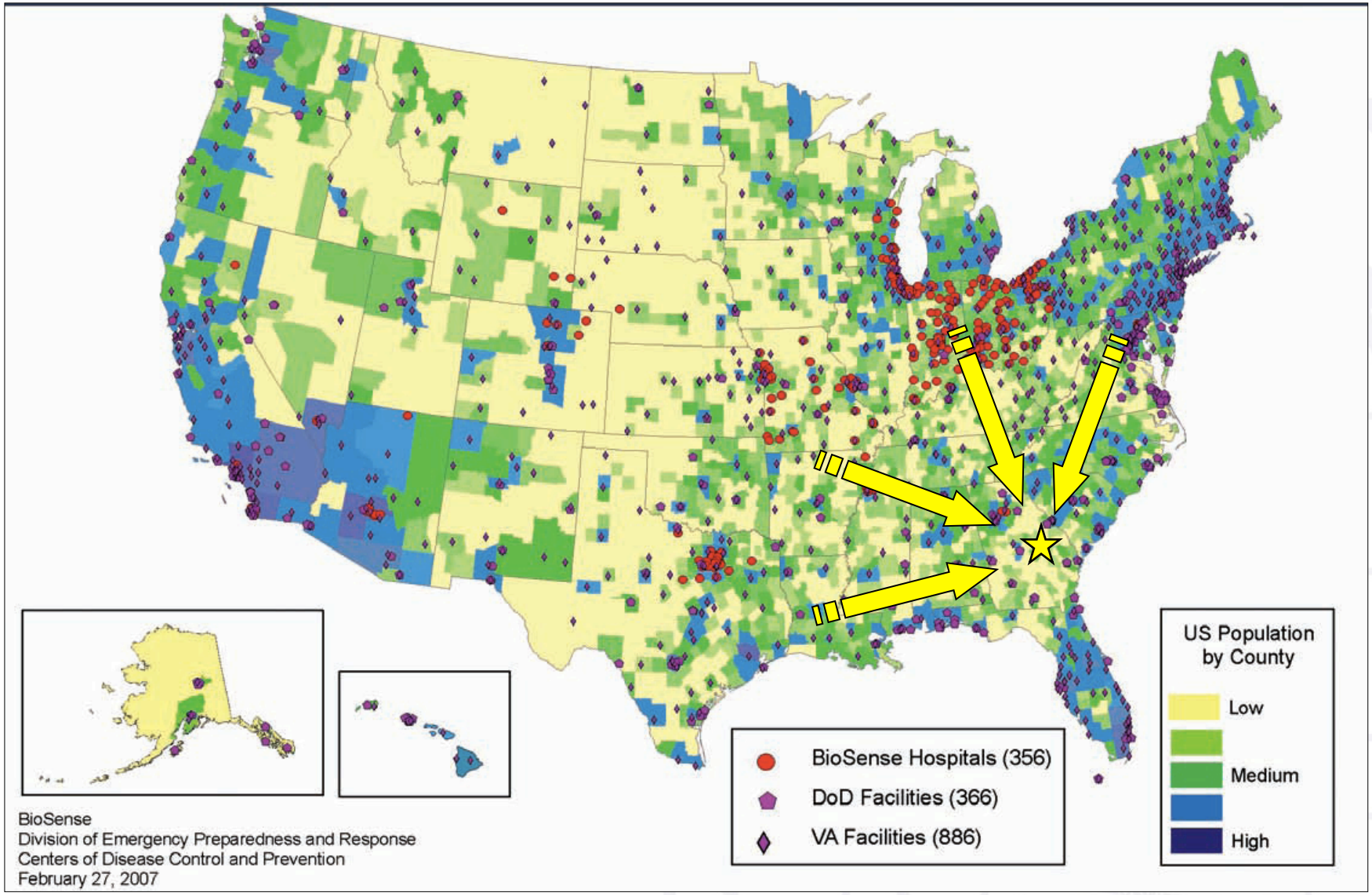
[1] www.whitehouse.gov/news/releases/2007/10/20071018-10.html
[2] CDC (www.cdc.gov/epo/dphsi/syndromic.htm, accessed 5/29/07)

Early Event Detection and Health Situational Awareness

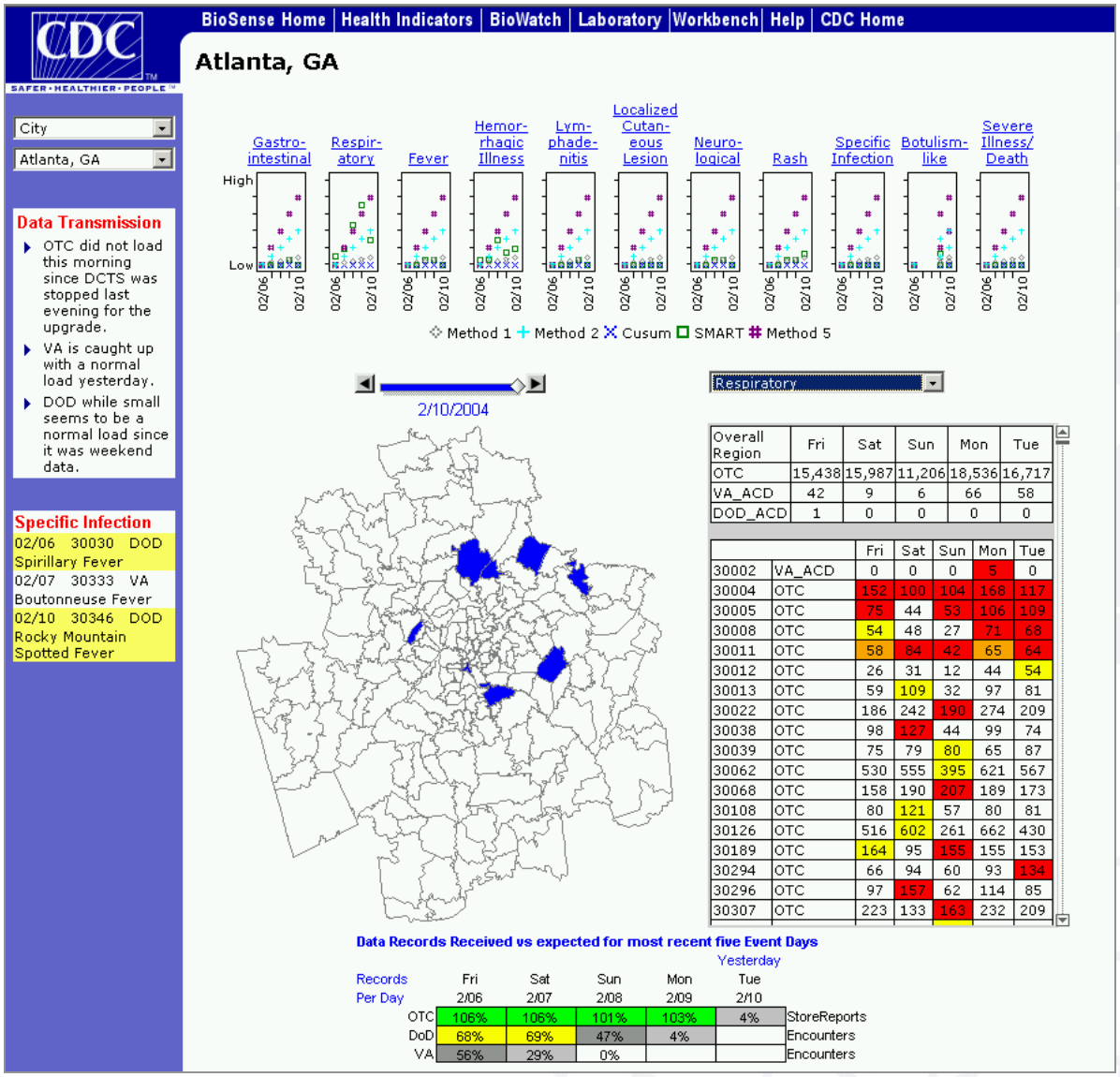
- "*Early Event Detection* (EED) is the ability to detect at the earliest possible time events that may signal a public health emergency. EED is comprised of case and suspect case reporting along with statistical analysis of health-related data. Both real-time streaming of data from clinical care facilities as well as batched data with a short time delay are used to support EED efforts." [1]
- "*Health Situational Awareness* is the ability to utilize detailed, real-time health data to confirm, refute and to provide an effective response to the existence of an outbreak. It also is used to monitor an outbreak's magnitude, geography, rate of change and life cycle." [1]

[1] CDC (http://www.cdc.gov/BioSense/publichealth.htm, accessed 10/11/08)

An Existing System: BioSense



BioSense Output



Latest Entry: Google Flu Trends

google.org Flu Trends

Google.org home

Flu Trends

Home

How does this work?

FAQ

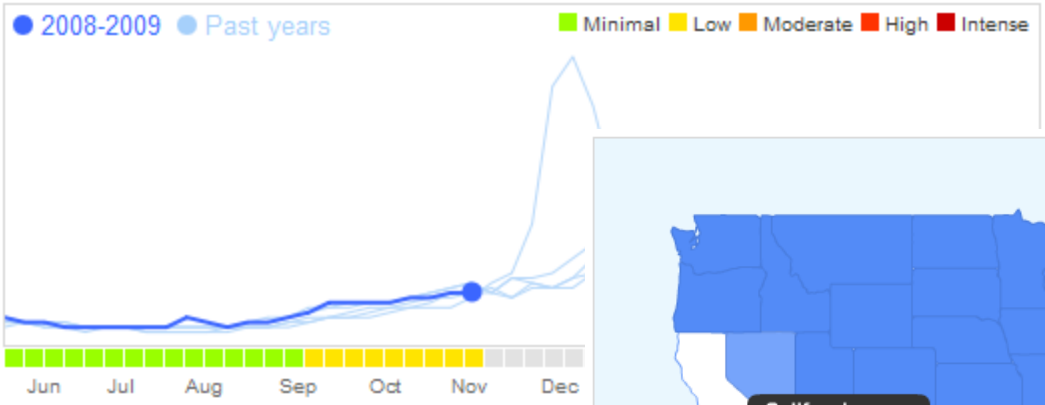
Download raw data

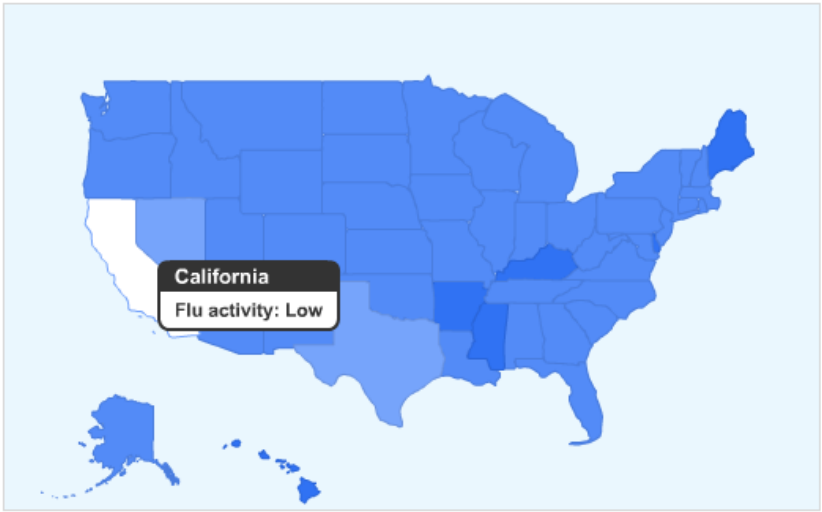
Explore flu trends across the U.S.

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity in your state up to two weeks faster than traditional systems. Read more »

United States flu activity: Low

Entire United States

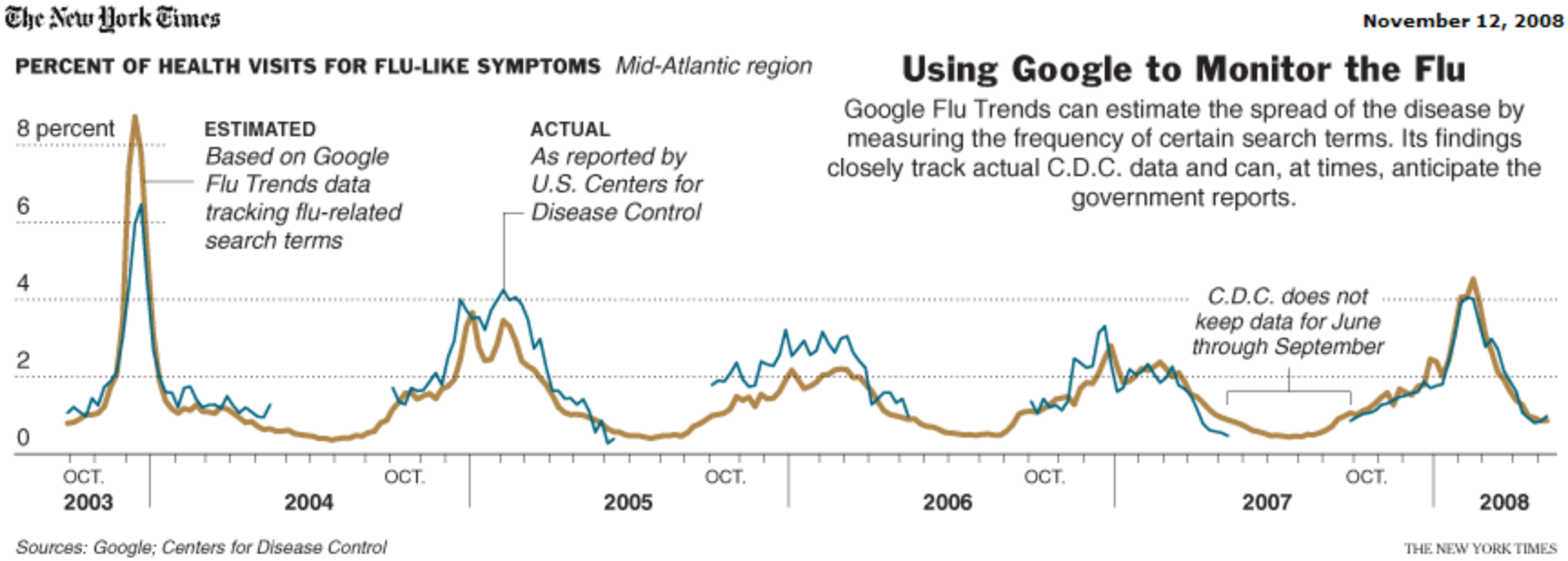




Data current through: November 10, 2008

See www.google.org/flutrends/

How Good is Google Flu Trends?

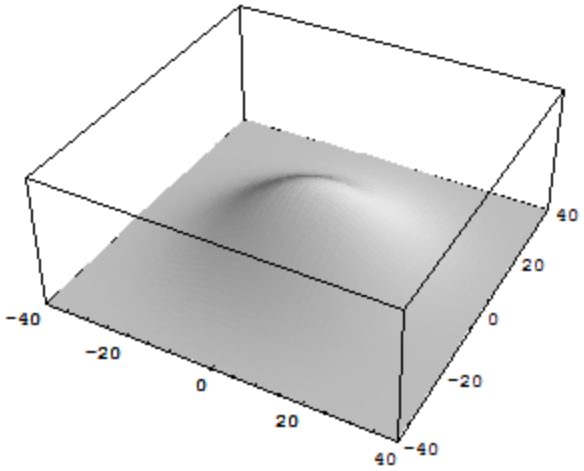


- Google search results correspond to CDC "sentinel physician" data
- Google says it was able to accurately estimate flu levels 1-2 weeks faster than published CDC reports

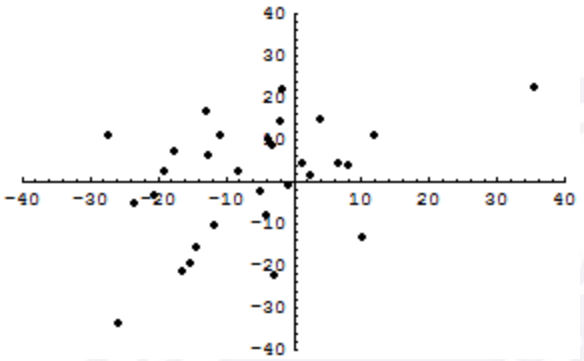
Research Goal

- Goal: Develop a method to identify and track changes in (local) disease patterns incorporating data in (near) real time
  - Is an outbreak/attack likely occurring?
  - If so, where and how is it spreading?
- Most methods focus on EED with aggregated (i.e., daily count) data
- Most common spatial method looks for clusters of cases

Illustrative Example



(Unobservable) spatial distribution of disease



Observed distribution of ER patients' locations

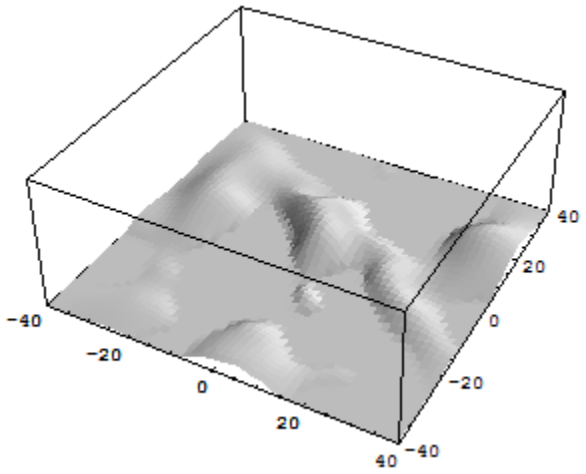
- ER patients come from surrounding area
  - On average, 30 per day
    - More likely from closer distances
  - Outbreak occurs at (20,20)
    - Number of patients increase linearly by day after outbreak

A Couple of Major Assumptions

- Can geographically locate individuals in a *medically meaningful* way
  - Data not currently available
  - Non-trivial problem
- Data is reported in a timely and consistent manner
  - Public health community working this problem, but not solved yet
- Assuming the above problems away…

Idea: Look at Differences in Kernel Density Estimates

- Construct kernel density estimate (KDE) of "normal" disease incidence using N historical observations
- Compare to KDE of most recent $w+1$ obs



But how to know when to signal?

Solution: Repeated Two-Sample Rank (RTR) Procedure

- Sequential hypothesis test of estimated density heights
- Compare estimated density heights of recent data against heights of set of historical data
  - Single density estimated via KDE on *combined* data
- If no change, heights uniformly distributed
  - Use nonparametric test to assess

Data & Notation

- Let $\mathbf{X}_i = \{X_{1i}, X_{2i}\}$ be a sequence of bivariate observations
  - E.g., latitude and longitude of a case
- Assume a historical sequence $\mathbf{X}_{1-N}, \ldots, \mathbf{X}_0$ is available
  - Distributed iid according to f_0
- Followed by $\mathbf{X}_1, \mathbf{X}_2, \ldots$ which may change from f_0 to f_1 at any time
- Densities f_0 and f_1 unknown

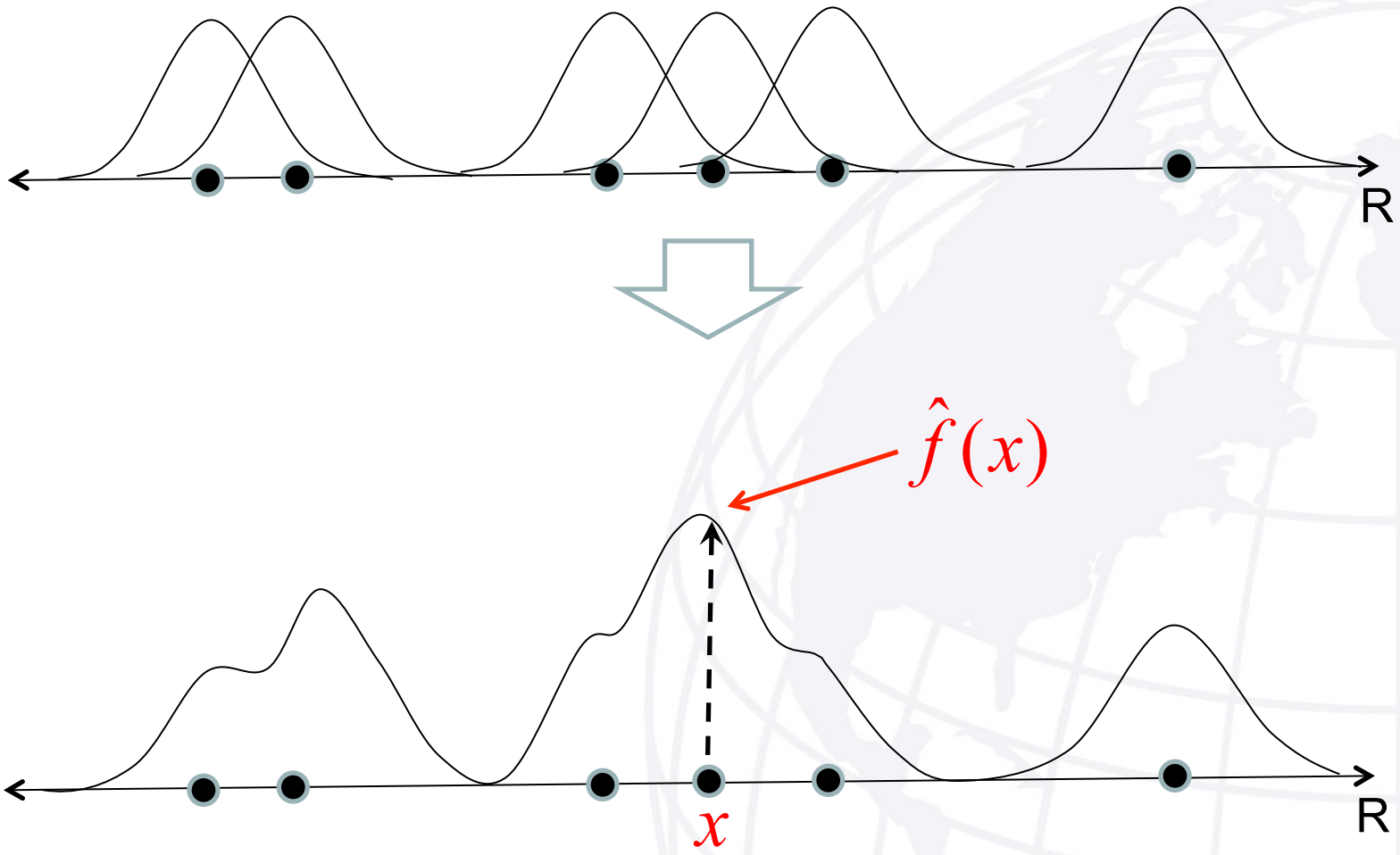
Estimating the Density

- Consider the $w+1$ most recent data points
- At each time period estimate the density

$$\hat{f}_n(\mathbf{x}) = \begin{cases} \dfrac{1}{N+n} \displaystyle\sum_{i=1-N}^{n} k_h\left(\mathbf{x}, \mathbf{X}_i\right), & n < w+1 \\ \dfrac{1}{N+w+1} \displaystyle\sum_{i=n-w-N-1}^{n} k_h\left(\mathbf{x}, \mathbf{X}_i\right), & n \geq w+1 \end{cases}$$

where k is a kernel function on $\mathbb{R}^2$ with bandwidth set to $h_i = \sigma_i \left(1/\left(N+w+1\right)\right)^{1/6}$

Illustrating Kernel Density Estimation (in one dimension)



Calculating Density Heights

- The density estimate is evaluated at each historical and new point
  - For $n < w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{1-N}), \ldots, \hat{f}_n(\mathbf{X}_0)}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_1), \ldots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$

  - For $n \geq w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{n-w-N-1}), \ldots, \hat{f}_n(\mathbf{X}_{n-w-1})}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_{n-w}), \ldots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$

Under the Null, Estimated Density Heights are Exchangeable

- *Theorem*: If $X_i \sim F_0$, $i \leq n$, the RTR is asymptotically distribution free
  - I.e., the estimated density heights are exchangeable, so all rankings equally likely
  - Proof: See Fricker and Chang (2008)
- Means can do a hypothesis test on the ranks each time an observation arrives
  - Signal change in distribution first time test rejects

Comparing Distributions of Heights

- Compute empirical distributions of the two sets of estimated heights:

$$\hat{J}_n(z) = \frac{1}{w+1} \sum_{i=n-w}^{n} I\left\{\hat{f}_n(\mathbf{X}_i) \le z\right\},$$

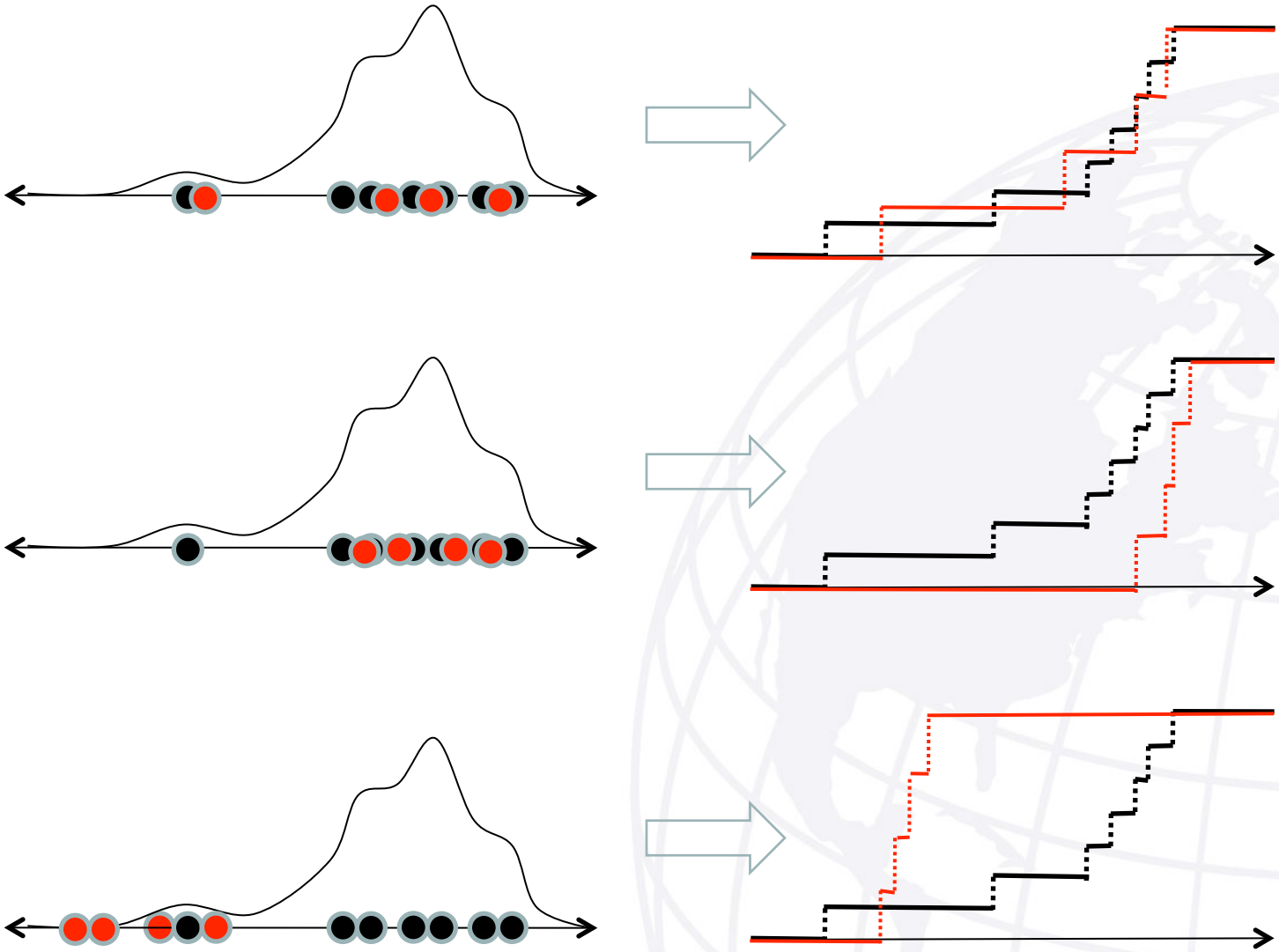
$$\hat{H}_n(z) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} I\left\{\hat{f}_n(\mathbf{X}_i) \le z\right\}$$

- Use Kolmogorov-Smirnov test to assess:

$$S_n = \max_z \left|\hat{J}_n(z) - \hat{H}_n(z)\right|$$

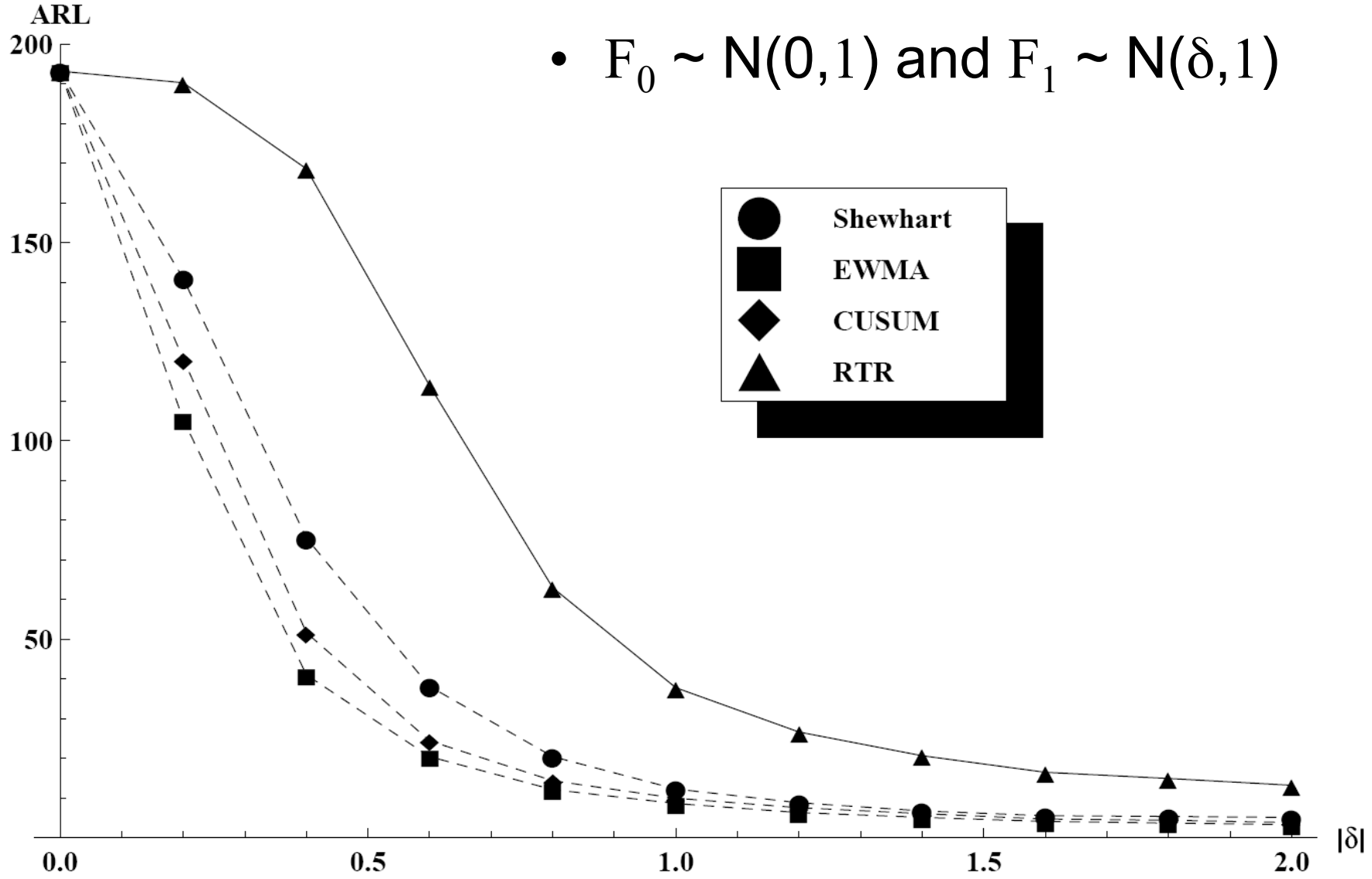
  – Signal at time $t = \min\left\{n : S_n > c\right\}$

Illustrating Changes in Distributions (again, in one dimension)



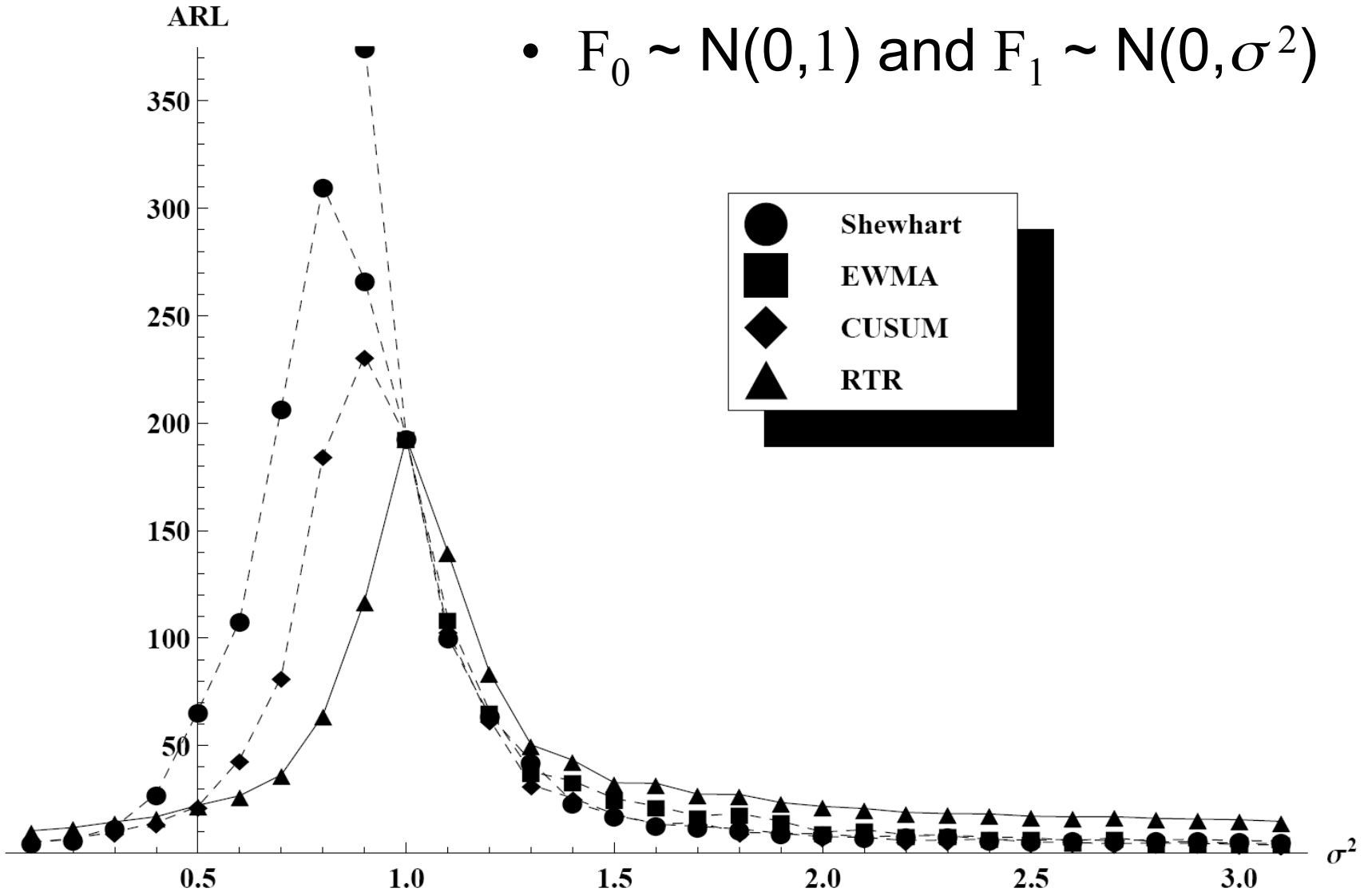
Performance Comparison #1

- $F_0 \sim N(0,1)$ and $F_1 \sim N(\delta,1)$



Performance Comparison #2

- $F_0 \sim N(0,1)$ and $F_1 \sim N(0,\sigma^2)$



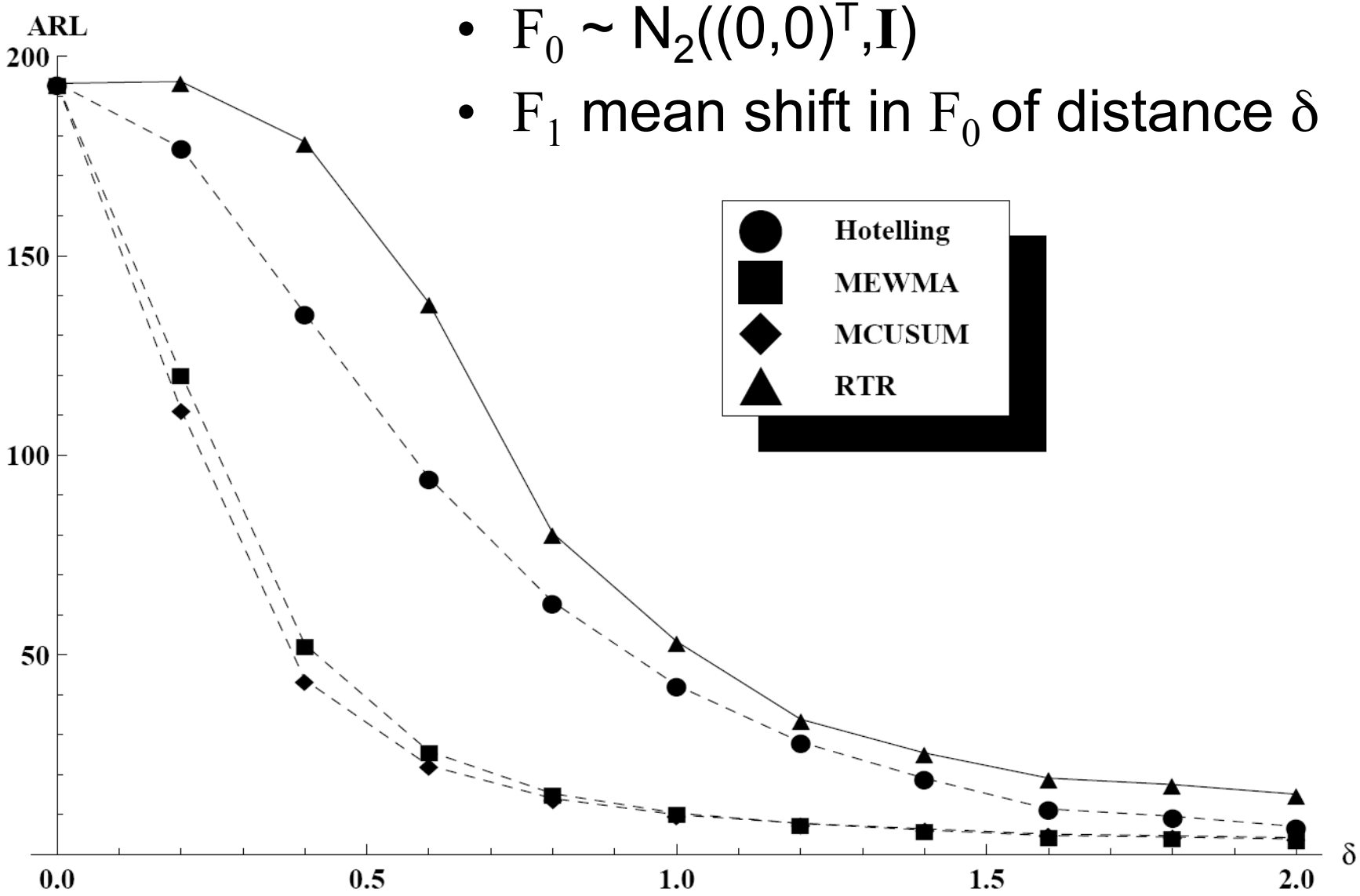
Performance Comparison #3

- $F_0 \sim N(0,1)$
- $F_1 \sim$ 

Distribution #	0	1	2	3	4	5	6	7	8	9	10
μ	0.0	0.435	0.6	0.715	0.8	0.866	0.917	0.954	0.980	0.995	1.0
σ^2	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	~ 0

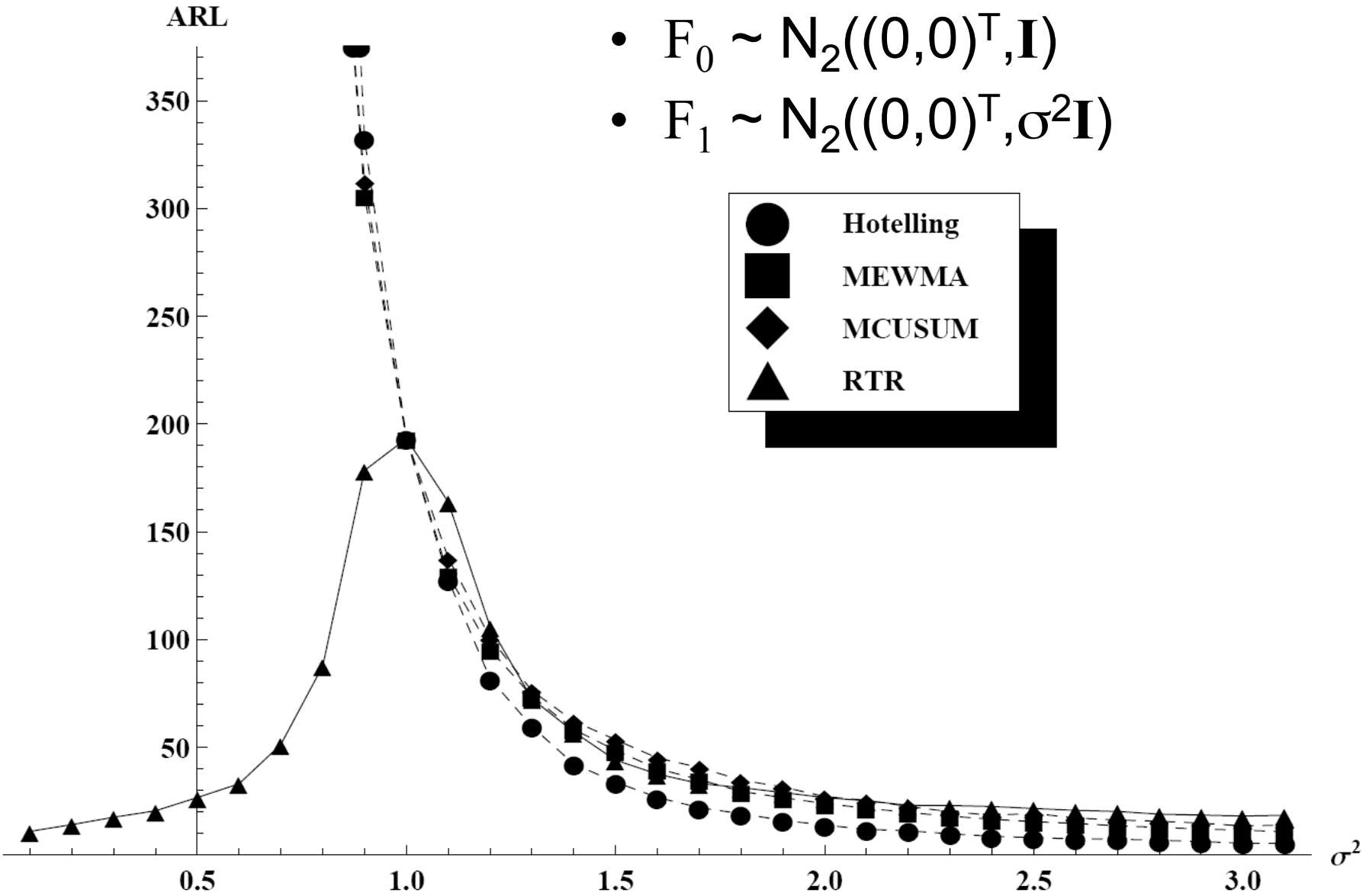
Performance Comparison #4

- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- F_1 mean shift in F_0 of distance δ



Performance Comparison #5

- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- $F_1 \sim N_2((0,0)^T, \sigma^2\mathbf{I})$



Setting the Threshold for the RTR

- How to find c?
  - Use ARL approximation based on Poisson clumping heuristic[1]:

$$A \approx \left[\left(\frac{6.16c\,[c + 0.5/(w+1)]}{1 + (w+1)/N} \right) \exp\left\{ -2\left(c + \frac{1}{2(w+1)} \right)^2 \left(\frac{1}{w+1} + \frac{1}{N} \right)^{-1} \right\} \right]^{-1}$$

- Example: c=0.07754 with N=1,350 and $w+1$=250 gives A=900
  - If 30 observations per day, gives average time between (false) signals of 30 days

[1] For more detail, see Fricker, R.D., Jr., Nonparametric Control Charts for Multivariate Data, Doctoral Thesis, Yale University, 1997.

Plotting the Outbreak

- At signal, calculate optimal kernel density estimates and plot pointwise differences

$$\Delta_n(\mathbf{x}) = \hat{h}_n(\mathbf{x}) - \hat{g}_n(\mathbf{x})$$

where

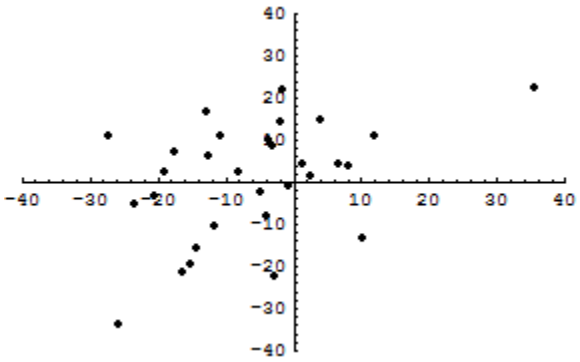
$$\hat{h}_n(\mathbf{x}) = \frac{1}{w+1} \sum_{i=n-w}^{n} k_h\left(\mathbf{x}, \mathbf{X}_i\right)$$

$$\hat{g}_n(\mathbf{x}) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} k_h\left(\mathbf{x}, \mathbf{X}_i\right)$$

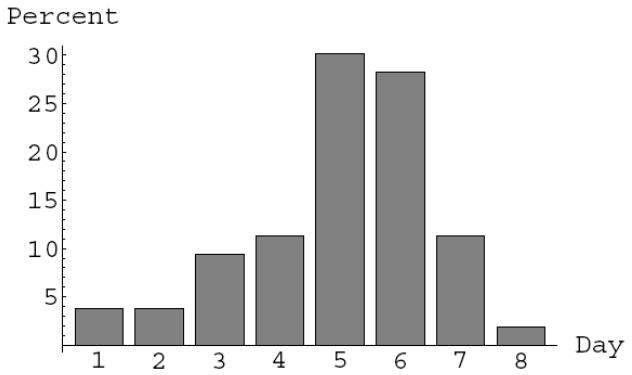
and $h_i = \sigma_i \left(\frac{1}{w+1}\right)^{1/6}$ or $h_i = \sigma_i \left(\frac{1}{N}\right)^{1/6}$

Example Results

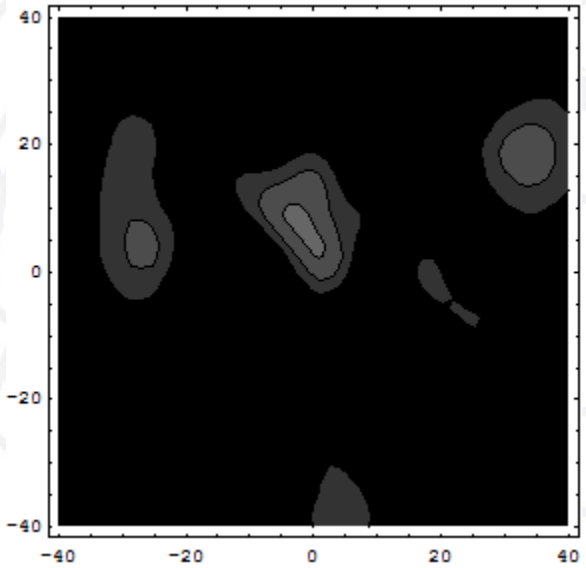
- Assess performance by simulating outbreak multiple times, record when RTR signals
  - Signaled middle of day 5 on average
  - By end of 5th day, 15 outbreak and 150 non-outbreak observations
  - From previous example:



Daily Data

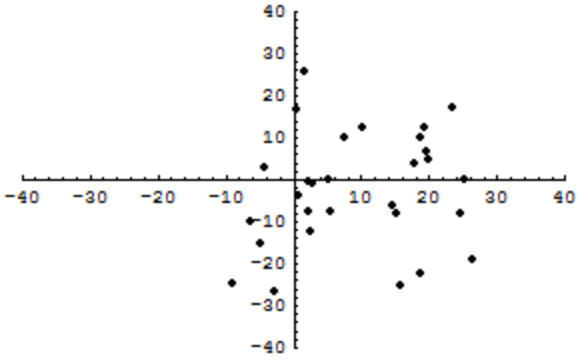


Distribution of Signal Day

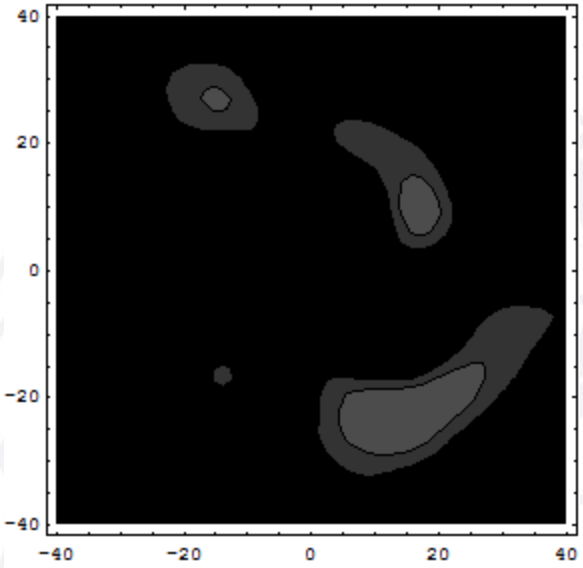


Outbreak Signaled on Day 7 (obs'n # 238)

Same Scenario, Another Sample



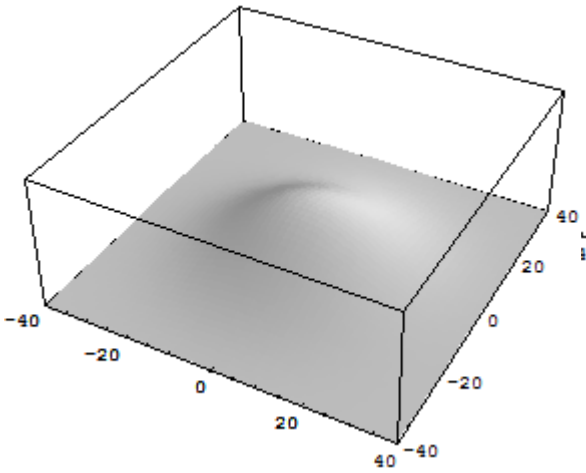
Daily Data



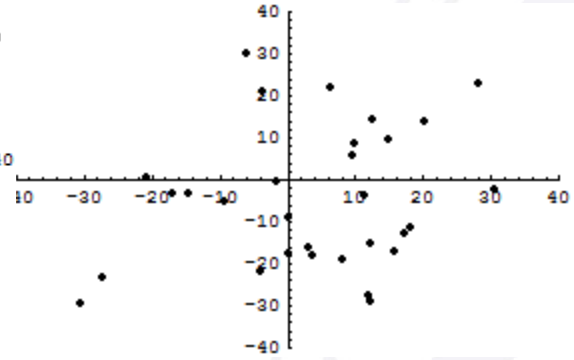
Outbreak Signaled on Day 5 (obs' n # 165)

Another Example

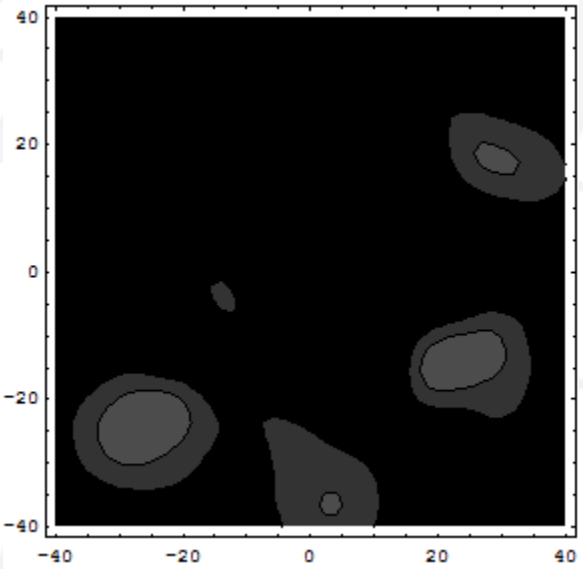
- Normal disease incidence ~ $N(\{0,0\}^t, \sigma^2\mathbf{I})$ with $\sigma=15$
  - Expected count of 30 per day
- Outbreak incidence ~ $N(\{20,20\}^t, 2.2d^2\mathbf{I})$
  - d is the day of outbreak
  - Expected count is $30+d^2$ per day



Unobserved outbreak distribution



Daily data

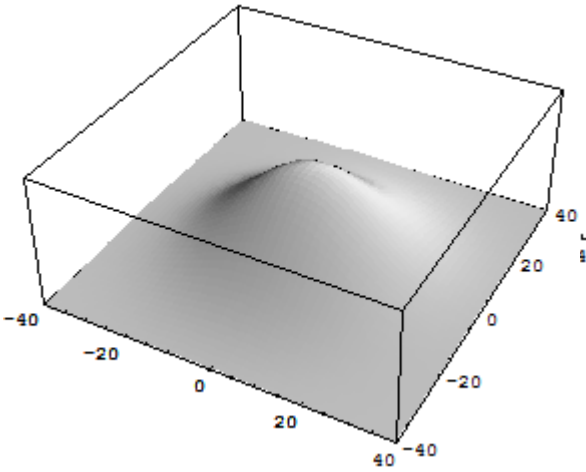


Outbreak signaled on day 1 (obs' n # 2)

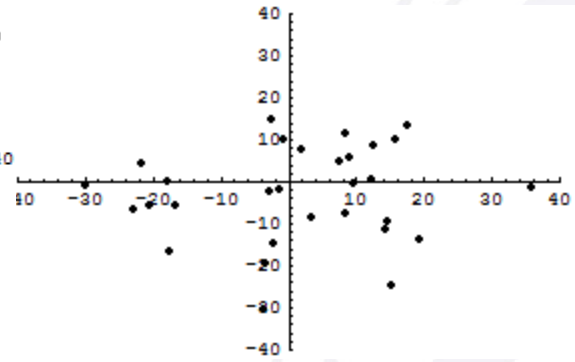
(On average, signaled on day 3-1/2)

And a Third Example

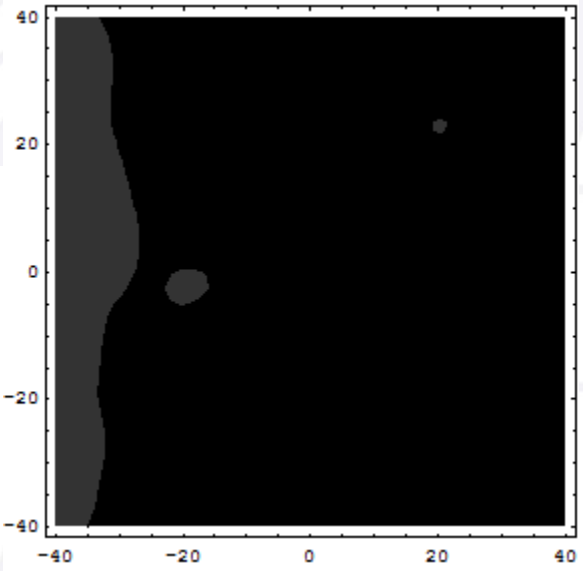
- Normal disease incidence ~ $N(\{0,0\}^t, \sigma^2 \mathbf{I})$ with σ=15
  - Expected count of 30 per day
- Outbreak sweeps across region from left to right
  - Expected count is 30+64 per day



Unobserved outbreak distribution



Daily data



Outbreak signaled on day 1 (obs'n # 11)

(On average, signaled 1/3 of way into day 1)

Advantages and Disadvantages

- Advantages
  - Methodology supports both biosurveillance goals: early event detection *and* situational awareness
  - Incorporates observations sequentially (singly) so can be used for real-time biosurveillance
    - Most other methods use aggregated data
- Disadvantage?
  - Can't distinguish increase distributed according to f_0
    - Won't detect an general increase in background disease incidence rate
      - E.g., Perhaps caused by an increase in population
      - In this case, advantage not to detect
    - Unlikely for bioterrorism attack?

Future Research

- Finish paper on RTR as general SPC methodology
- Looking to see if plotting $\left\{\mathbf{X}_i : \left|\hat{J}_n(z) - \hat{H}_n(z)\right| > c\right\}$ on the contour plots helps to show where the outbreak is occurring
- Compare the performance of the RTR for detecting outbreak clusters to commonly used methods
  - SatScan (Kulldorff)
  - SMART (Kleinman)

Selected References

Detection Algorithm Development and Assessment:

- Fricker, R.D., Jr., and J.T. Chang, The Repeated Two-Sample Rank Procedure: A Multivariate Nonparametric Individuals Control Chart (in draft).
- Fricker, R.D., Jr., and J.T. Chang, A Spatio-temporal Method for Real-time Biosurveillance, Quality Engineering, 20, pp. 465-477, 2008.
- Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu, Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, *Quality Engineering*, **20**, pp. 478-494, 2008.
- Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr., A One-Sided MEWMA Chart for Health Surveillance, *Quality and Reliability Engineering International,* **24**, pp. 503-519, 2008.
- Fricker, R.D., Jr., Hegler, B.L., and D.A Dunfee, Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms, *Statistics in Medicine,* **27**, pp. 3407-3429, 2008.
- Fricker, R.D., Jr., Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance,* **3**:1, 2007.

Biosurveillance System Optimization:

- Fricker, R.D., Jr., and D. Banschbach, Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods, in submission.

Background Information:

- Fricker, R.D., Jr., and H. Rolka, Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4-13, 2006
- Fricker, R.D., Jr., Syndromic Surveillance, in *Encyclopedia of Quantitative Risk Assessment*, Melnick, E., and Everitt, B (eds.), John Wiley & Sons Ltd, pp. 1743-1752, 2008.